'Making it count': incentives, student effort and performance

Article (Accepted Version)

"Making it count"

Incentives, Student Effort and Performance

Arnaud Chevalier
(Royal Holloway, University of London, and IZA)

Peter Dolton
(University of Sussex and CEP, LSE and IZA)

Melanie Lührmann
(Royal Holloway, University of London, IFS and MEA)

Abstract

This paper examines how incentives to participate in online assessments (quizzes) affect students' effort and performance. Our identification strategy exploits within-student weekly variation in incentives to attempt online quizzes. We find tournament incentives and participation incentives to be ineffective in increasing quiz participation. In contrast, making the quiz counts towards the final grade substantially increases participation. We find no evidence of displacement of effort between weeks. Using a natural experiment which provides variation in assessment weighting of the quizzes between two cohorts, we find that affected students obtain better exam grades. We estimate the return to 10% assessment weighting to be around 0.27 of a standard deviation in the in-term exam grade. We find no evidence that assessment weighting has unintended consequences, i.e., that increased quiz effort: displaces effort over the year; reduces other forms of effort; or reduces (effort and thus) performance in other courses. Finally, assessment weighting induced effort increases most for students at and below median ability, resulting in a reduction of the grade gap by 17%.

JEL Code: I23, D20
Keywords: Incentive, Effort displacement, Effort, Higher Education

Acknowledgement: We wish to acknowledge the help of the administrative staff at RHUL and thank Eleftherios Giovanis for research assistance. We are grateful to Macke Raymond, Edwin Leuven, Hessel Oosterbeek, Susan Dynarski, John Bound, Santiago Oliveros and seminar participants at CESifo, RES, Sussex, Michigan, City University, Nuremberg, SOFI Stockholm, NIESR, SFI and IZA for comments. The editor and anonymous referees also provided numerous valuable comments that improved this manuscript. This project was partially financed by the RHUL Faculty Initiative Fund.

Arnaud Chevalier, Royal Holloway, University of London, Economics, Egham TW20 0EX, UK. , arnaud.chevalier@rhul.ac.uk (Corresponding author). Chevalier is also affiliated to the Geary Institute, Dublin; SFI, Copenhagen, ROA, Maastricht, and the Life Course Centre, Queensland.

Peter Dolton, University of Sussex, Economics, Brighton, BN1 9RH, UK. Dolton is also affiliated to the Center for Economic Performance (CEP LSE), CESifo and IZA.

Melanie Lührmann, Royal Holloway, University of London, Economics, Egham TW20 0EX, UK. She is also affiliated to the Institute for Fiscal Studies (IFS) and the Munich Center for the Economics of Aging (MEA).

1. Introduction

In higher education, a sizeable fraction of students fail their courses, and a substantial minority drop out. This may be due to a lack of study effort, possibly driven by uncertainty about returns to study effort, high discounting of the future, or (mis)perception of their own ability. If intrinsic motivation does not suffice to induce satisfactory effort, then what interventions might help to increase student effort and performance? A common tool to engage students with their studies is to provide online computer learning assessments in the form of a "quiz." This study evaluates how incentives affect quiz effort. We first compare the efficiency of different incentives in inducing a student to provide effort, exploiting weekly incentive variation for all students. To evaluate longer-run effects of quiz participation on learning, we rely on a natural experiment where two cohorts were subjected to different sets of quiz incentives. In particular, the control cohort was offered quizzes, but these were not part of their course assessment. In contrast, the second (treated) cohort were rewarded with assessment weightings from grades obtained in some pre-determined weekly quizzes which counted towards the course's final grade. To facilitate a comparison between the treated and control groups we use propensity score matching to provide evidence that the two cohorts do not statistically differ on observable and usually unobserved characteristics. We can then assess how assessment rules affect course performance.

The data pertain to two cohorts of first year undergraduate economics students at a large college of the University of London (for which we have records of participation in weekly online quizzes for the course we investigate). The quizzes are designed to foster continuous learning so that quiz participation may help increase students' learning and performance. In practice, the simple provision of these educational resources does not suffice to engage students; quiz participation in non-incentivized weeks is only 29%. However, students were exposed to varying incentives (or lack thereof) across weeks, and we estimate their effort

responses accordingly. In particular, we investigate the relative effects of: i) a participation incentive, i.e., the provision of additional study material conditional on quiz participation, ii) a tournament incentive in the form of a book voucher for the best quiz performance of the week, and iii) assessment weighting, whereby the grade obtained in the quiz counts towards the course grade. We find that the participation and tournament incentives are ineffective in raising quiz effort. In contrast, however, even small assessment weights of 2.5% to 5% of the final grade have a large impact; they increase weekly quiz participation by 42 (respectively 62) percentage points. This compares with a 73 percentage point increase in participation when the quiz is made compulsory.

Effort responses to incentives differ by ability as measured via university entry grades. Lower ability students are less likely to exert effort in the absence of incentives, potentially due to lower intrinsic motivation to study. Frey and Jegen (2002) suggest that extrinsic motivation, e.g., incentives, can counteract the lack of intrinsic motivation, but they also warn of the opposite effect when intrinsic motivation is high. We do not find detrimental effects of assessment weighting on effort among high ability students (who are more likely to participate in quizzes in the absence of incentives). However, it induces relatively higher quiz effort among lower and median ability students, in line with extrinsic motivation inducing this group to study harder. As a result it would seem that assessment weighting helps to "level the playing field."

Assessment weighting may contemporaneously increase effort in incentivized weeks but could *discourage* effort in non-incentivized weeks. We investigate this effort 'displacement hypothesis', which could reduce the effectiveness of incentives inter-temporally, using two complementary approaches: first, we estimate dynamic models of effort choice using leads and lags of the incentive treatments; secondly, we use the first week of the year in which students did not yet know about the incentive scheme as a benchmark. Both approaches yield the same result: we find no displacement, i.e., there is no shifting of effort during the term, and

assessment weighting unambiguously increases total quiz effort over the term. Interestingly in the first week of the year, and in other non-incentivized weeks, the quiz participation rates are close between the two cohorts, highlighting their similarity.

In the second part of the paper, we investigate whether rewarding effort in some weeks does indeed increase exam performance. To do so, we exploit a natural experiment; Cohort 2, our treatment cohort, was subjected to six assessment weighted quizzes instead of two compulsory ones; other participation incentives remained the same. Despite compulsory quizzes having a larger effect on weekly participation, the treated students completed 1.3 additional quizzes (out of a 7 quizzes available before the in-term test) per term, on average. We then use propensity score matching methods to account for cohort composition differences and exploit the exogenous variation in quiz assessment to identify its impact on student performance. We find that a 10% assessment weighting, split across three quizzes with smaller weights, increases exam performance substantially by 0.27 of a standard deviation. Two mechanisms are potentially at play; for the treated cohort, the frequency of incentivized quizzes was greater; second, effort in assessed quizzes might be more productive than in unweighted compulsory quizzes, as such more learning took place during quizzes for the treated cohort.

This reduced form effect could be driven by unobserved differences between the two cohorts and not by changes in the incentives faced by students during the term. However, we note that in the first week of the year, the quiz participation rates of the two cohorts were similar, so they do not appear to differ in intrinsic motivation. Furthermore, we test that the performance improvement is driven by the increase in quiz participation and not directly by a cohort effect nor by changes in the productivity of quiz; i.e. the grade effects are driven by the increased quiz completion and not by any other differences between the two cohorts. The performance improvement is (quantitatively and qualitatively) consistent with our estimates of the effort response to assessment weighting, and the estimated grade returns to additional quiz

effort. The latter amounts to a grade increase by 0.165 of a standard deviation per additional completed quiz. The effort return is nonlinear and concentrated around a completion of at least 4 out of 7 quizzes. Compared to the initial regime, assessment weighting increases the proportion of students completing at least 4 tests per term by 40%. We also provide evidence of longer-run effects of assessment weighting and find similar size effects (if not always precisely estimated) on final exam grades, and on the overall course grade in the treated course, in other first year modules and in related second year courses.

Insights from behavioral economics suggest that incentives (nudges) can increase educational effort by providing students with non-financial returns to their input or a better study structure (see reviews by Lavecchia *et al.* (2016), Gneezy, Meier and Rey-Biel (2011) and Koch, Nafziger and Skyt-Nielsen (2015)). As such, various studies have considered assessment weights in higher education. Pozo and Stull (2006) run an experiment where an initial math assessment counted (or not) towards the final grade; they report greater levels of effort on that assessment among incentivized students and an improvement in the overall score by 2 percentage points, especially for weaker students. Their estimated effect is the joint impact of the provision of math training and assessment weighting, which may both improve performance separately. Closer to our study, other papers have estimated the effect of compulsory/assessed homework on students' exam performance. Grodner and Rupp (2013) randomize compulsion to homework within a cohort, and report improvement in test scores by 6%. Trost and Salehi-Isfahani (2012) also randomize compulsory participation to homework within a section in different weeks. For three topics, students were randomly assigned to a voluntary or compulsory assessment group. Students in the compulsory group perform better in the short-run, but this effect disappears by the end of semester.

Rather than compulsion, Emerson and Menken (2011) randomize, within section, assessment weighting of homework representing up to 18% of final grade. The treated group

outperforms control group students by 4% on average - but only in tests directly covered by the assessed homework (Emerson and Menken, 2011) – and not the Test of Understanding in College Economics. Similarly, Grove and Wasserman (2006) rely on a natural experiment, leading to differences in assessment weighting for homework (15% towards the final grade) within a cohort, and report a final grade improvement of 2 percentage points. Overall, the literature has provided evidence that in the short-run both compulsion and assessment weighting can be effective at increasing grades.

We provide additional evidence in favor of the effectiveness of assessment weighting in raising effort and increasing performance, extending the literature in four important ways: first, we test the effect of several incentives within the same student population and find assessment weighting to be most effective if compared to prizes and participation incentives. Indeed, we find that a 5% assessment weight generates an increase in that week quiz participation that is 86% of the increase generated by compulsion. Second, since students face different incentives week by week, we can control for students' unobserved characteristics in our analysis of the effort impact of incentives. Third, we investigate heterogeneity effects with respect to gender and pre-determined ability and include usually unobserved characteristics such as risk attitudes. We find that effort increases are highest among lower and median ability students: since these are students with low observed effort in the absence of incentives, extrinsic motivation through incentives activates this group's effort. At the same time, we find no evidence that incentives crowd out intrinsic motivation to study among those with initially high effort (who also tend to be high ability students).

Our fourth and main contribution is to show in detail that inducing effort via assessment weighting does not displace effort elsewhere—neither inter-temporally nor inter-subject. Final grades, pass rates and grades in other modules were not negatively affected by the additional effort provided on the treated module during term time. In contrast, our results point to

increases in overall effort throughout the course, and potentially positive spillover effects on performance in other courses in the same year, and related courses in the second year. We thus conclude that frequent assessment weighting has unequivocally positive effects on course effort, and even results in better academic performance than less frequent compulsory quizzes. Grades do not only increase at the mean; performance increases are concentrated around the median of the grade distribution, narrowing the performance gap.

Our results also contribute to the broader literature which considers effective incentive mechanisms in the context of educational performance. Fryer (2011, 2013) concludes from a series of experiments in schools that incentives rewarding inputs tend to be more successful at generating grade improvements than incentives targeting outputs. Financial incentives, which tend to reward output, tend to yield small performance improvements—also in studies focusing on higher education (see Lavecchia, Liu and Oreopoulos (2016) for a review). These moderate effects may be a consequence of the crowding out of intrinsic motivation (Frey and Jegen, 2002) or a mismatch between achievement targets and students' ability (Camerer and Hogarth, 1999). The magnitude of the performance effect in our study is similar to those achieved through large financial incentives awarded for increased grades (see Angrist *et al.*, 2009, 2014, Leuven *et al.* 2010, Cha and Patel 2010, and Barrow and Rouse 2013). However, our effects are achieved at much lower overall cost. Additionally, financial incentives may be difficult to scale up considering the limited resources of higher education institutions. Our results are consistent with the hypothesis that rewarding inputs leads to increased performance, and this can be achieved using a cheap and easy-to-implement incentive mechanism: assessment weights. Of course, compulsion has an even larger effect on quiz participation but there may be reasons to prefer assessed tests, which we have not tested in this paper. Compulsion, while good at increasing quiz participation may not lead to as much study effort as assessment, since they are no rewards to effort, thus leading to a smaller continuous learning effect. Assessed

tests, by reducing the weight on final exams, decrease performance volatility as a "bad day" will have less impact on the final grade in a course with frequent assessments that carry small weights, which anecdotal evidence suggests is disliked by students. Finally, compulsion also leads to administrative burden to assess the validity of extenuating circumstances of students who failed to comply with the compulsion. This administrative burden increases linearly in the amount of compulsory tests administered.

In the remainder of the paper, we first describe the institutional set-up and the data and identification strategy (Section 2). Section 3 then presents empirical results, and Section 4 concludes.

2. Study design and Data

The study was conducted on the course "Principles of Economics" (henceforth: "Principles"), which is compulsory for all students enrolled in economics at a large college of the University of London. There is no self-selection into the course or any of its delivery components (i.e., the lecture or seminar). "Principles" represents a quarter of students' first year undergraduate curriculum load. It is a "high stakes" module which students need to pass in order to progress to the second year. "Principles" runs over two terms (Fall and Spring) for a total of 20 weeks. Each week, students attend a two-hour lecture and a (compulsory) one-hour small group seminar. All course materials are available online via a dedicated interface, which also hosts a set of on-line multiple choice quizzes testing the understanding of the concepts taught in the previous week.

The two course instructors (each teaching for one term), contents, materials, delivery, lecture times, communication and the quiz question database remain identical across the two years under study. In both years, students were informed about quizzes in the same way and encouraged to participate. After each lecture, students could complete quizzes within a

predefined window of one to three days. After the due date they received information on their overall score, their answers, the correct solution to each question, and in some cases detailed explanations. For assessed quizzes, students were given a 60 minute completion window within a specified 24-hour period. For each student, questions were randomly drawn from a large question bank.

In a given week, all students faced the same incentive to complete the quiz. However, these incentives varied across weeks; Table 1 reports the timing across weeks. Incentives were repeated at least twice throughout the year and their timing across terms differs so that they are unlikely to capture week-specific effects. This "within-student" design allows us to account for students' unobserved characteristics.

The first incentive is a simple *participation incentive* where students gain access to detailed seminar exercise solutions if they participate in the weekly quiz ("Solution"), conditional on achieving at least a very low quiz grade of 30%. The second is a *performance incentive* in the form of a £20 book voucher for the best quiz performance ("Voucher"), whose winner is announced in class/by email. We expect this tournament to increase quiz participation and performance among those students who believe they have a chance of winning, i.e., those with higher ability (or those who are overconfident). At the same time, a tournament setting may have detrimental effects on others via a discouragement effect (Cason *et al.*, 2010).

Additionally, we exploit a natural experiment in which, for one cohort (cohort 2), in each term three quizzes carried assessment weighting: two out of the weekly online quizzes contribute 2.5% ("Assessed 2.5%") and one counts 5% ("Assessed 5%") towards the final course grade, amounting to a total of 10% per term, and 20% for the full year. The final exam accounts for the remaining 80%, while for Cohort 1, it accounts for 100% of the final course grade. If these assessment weights are salient enough, we expect an increase in student participation (and performance) in assessed quizzes relative to non-assessed ones. For the first cohort, only two

quizzes were mandatory ("Compulsory"). Students were informed that participation would be a prerequisite for entry to take the final exam. This threat was not strictly enforced, but we expect it to be credible among first year students. In several weeks, quizzes were not incentivized, so that only the desire to obtain feedback or intrinsic motivation induced students to participate.

Our data consists of the electronic records of quiz participation and in-term tests for two consecutive cohorts of students. We match it with administrative data containing information on final exam performance, students' characteristics, and ability (school completion grades). We further obtain a set of measures of preference parameters from a questionnaire survey, which has been conducted for several years in the first week of lectures. Excluding repeating students and students that drop out during the year, Cohort 1 consisted of 206 students, and Cohort 2 of 242 students. More information on our sample is provided in Appendix B.

In our analysis, we first estimate the effect of incentives on weekly quiz participation (q). Specifically, we compare students' responses to different types of incentives to participate in the weekly quiz. To capture unobserved heterogeneity, we specify the following model with individual fixed effects, μ_i:

$$q_{it} = \propto + \sum_z \beta_Z \, Incentive_{zt} + \sum_k \delta_k \, x_{ikt} + \; \varphi T_t + \mu_i + \varepsilon_{it}, \qquad (1)$$

where the subscripts refer to individual i at time t (measured in weeks). Each individual is exposed to a set of z incentives that are week-specific. All time-invariant student characteristics, including pre-determined ability, motivation, or work ethic, are absorbed in the individual fixed effect (μ_i). The x_{ikt} term conditions for the k measures of weekly variation in academic burden, such as assignment deadlines in other courses. We allow for time fixed effects T in the form of a term dummy and a term-specific linear trend in week. Standard errors are clustered at the week and cohort level, i.e., within a given incentive treatment, so as to account for any unobserved correlation between observations in a given week.

In addition to estimating the impact of incentivized quizzes relative to non-incentivized ones, we estimate additional specifications to test for potential displacement of effort between incentivized and non-incentivized weeks by using only the first week of the year as the control condition. Students were only informed about the weekly incentives after the first week and may treat all other non-incentivized weeks as potentially affected by the treatment. This allows us to test for displacement of effort between incentivized and non-incentivized weeks.

To assess possible displacement further, we additionally estimate model (1) including lags (leads) of incentives:

$$q_{it} = \propto + \sum_z \beta_Z \, Incentive_{zt} + \sum_z \gamma_Z \, L(Incentive_{zt}) + \sum_k \delta_k \, x_{ikt} + \; \varphi T_t + \mu_i + \varepsilon_{it} \, , \quad (2)$$

where L is a lag (lead) operator; i.e., we assess whether the last week's or the next week's incentive has an effect on this week's participation behavior.

Next, we sum up the number of submitted weekly quizzes before the in-term exam ($Q = \sum q$) to capture the effect of assessment weighting on overall quiz effort. We obtain two overall effort measures for each student, one for each term. We then estimate the effect of assessment weighting on overall quiz participation using the following specification:

$$Q_{ip} = \propto_Q + \theta_Q C_i + \sum_k \delta_{Qk} \, x_{ik} + \gamma_Q P + \varepsilon_{Qip} \, , \quad (3)$$

where C_i is a dummy that takes the value 1 if individual *i* is exposed to assessment weighting incentives, and P is a term dummy.

Of course, quiz participation is only of interest if it has a long-term effect on exam performance. A similar reduced-form model to (3) is used to estimate the effect of assessment weight incentives on student performance, according to the following model:

$$S_{ip} = \propto_S + \theta_S C_i \; + \sum_k \delta_{Sk} \, x_{ik} + \gamma_S P + \varepsilon_{Sip} \, , \quad (4)$$

where S_{ip} denotes the in-term grade of individual *i* in term P. Our identification strategy relies on differences in the incentives to provide effort between the two groups. Our estimates would

be biased if there were composition differences between the two cohorts. Columns 1 and 2 in Table 2 show some group differences. On average, Cohort 2 is about 3 months younger; has a larger fraction of males (64% relative to 53%); more economics majors; is of slightly lower academic ability - as measured by entry grades - and contains more British students. These compositional differences may have been due to higher enrollment rates in cohort 2 in anticipation of an impending reform in education financing which increased tuition fees (for subsequent cohorts). These compositional differences may induce a sample selection resulting in a downward bias in our estimation, as Cohort 2, which is affected by the assessment incentive, is slightly academically weaker at entry into university. To overcome this problem, we rebalance our sample using propensity score matching. We match based on age, gender, nationality, degree major and ability. 94% of Cohort 2 individuals are matched, highlighting the large amount of common support between the two groups. (see Appendix B). Columns 3 and 4 of Table 2 show that the matched cohorts are, as expected, balanced in terms of the observable characteristics used in the matching process (see Panel A). In Panel B, we show that the two cohorts are also balanced with regard to usually *unobserved* characteristics that were not part of the matching procedure. We rely here on survey measures of students' patience, risk attitude, and their self-confidence, which have been shown to be related to academic performance and a wide range of economic choices (for an overview, see Dohmen et al., 2010; for more detail, refer to Appendix A.1). We take this as further evidence for the quality of our matching approach; the two groups are now identical in terms of observable and (potentially) unobservable characteristics. For the remainder of the analysis, we reweight individuals in the control group (cohort 1) using their matching frequency.

To further alleviate concerns that the reduced form effects are driven by cohort differences that are unrelated to cohort composition, we also estimate the direct effect of quiz participation on exam performance, accounting for cohort effect. Moreover, we introduce an interaction

between the number of quizzes completed and cohort to test whether the effect of quiz participation on exam grades differ between cohorts.

$$S_{ip} = \propto_S + \vartheta_S Q_{ip} + \theta_S C_i + \pi_S Q_{ip} C_i + \sum_k \delta_k x_{ik} + \gamma P + \varepsilon_{ip} , \quad (5)$$

For the reduced form effects presented in (4) to be interpreted in a causal way, we would need the estimates in (5) on the direct cohort effect (θ_S) and the interaction term (π_S) to be insignificant; i.e. the effects on grades are driven only by quiz participation, not by differences between cohorts.

Panel C of Table 2 contains descriptive measures of the change in effort and performance across the matched groups. Students in Cohort 2 are significantly more active in continuous learning via quizzes: They participate in 23% more quizzes than Cohort 1 and their quiz grades are 10% higher. Even the time spent on each quiz increases significantly for this group. Term-time exam grades are 3.5 points or 5% higher, and final exam grades are up by 3.3 points, or 6.5%.

3. Results

3.1 Incentives and Weekly Effort

In this section, we investigate the effectiveness of different incentives for students to engage in continuous learning. Quiz participation rates in week 1, our baseline week before any treatment announcement, is indistinguishable between the treatment and control group (50% vs. 53%, two-sided, t-value=0.52), suggesting that the intrinsic motivation of the two cohorts is similar at the onset of the academic year. The low quiz participation in weeks without incentives suggests that students' demand for feedback is low when obtaining such feedback requires unrewarded effort (see Figure 1). Both cohorts allocate their effort similarly in the absence of incentives, as their participation in weeks without incentives is almost identical.

"Soft" incentives—book vouchers and the provision of seminar solutions, marked by dotted vertical lines in Figure 1—do not appear to have much impact on the quiz participation of either cohort. In contrast, quiz effort differs markedly in weeks with assessment weights for the second cohort (marked by dashed (5% weighting) and solid vertical lines (2.5% weighting)) or for the first cohort only, compulsion (dashed vertical lines). In weeks with assessment weights, participation in Cohort 2 spikes at above 70% and is substantially higher than participation in the same weeks in Cohort 1, particularly during the Spring term. These figures suggest a strong reaction to assessment incentives. Indeed, 2.5% weighting results in a participation rate that is 83% of the participation rate achieved in a compulsory assessment. These observations are confirmed in the econometric analysis (see Table 3). We estimate the fixed effects model from equation (1) relative to non-incentivized weeks to identify incentive effects on student effort. We use the matched sample and cluster standard errors at the cohort-week level.

The "Solution" incentive, which gives access to problem-set solutions conditional on quiz participation, has no statistically significant effect, possibly because students can share problem-set solutions. The £20 book voucher for the best quiz performance *reduces* participation marginally by 9 percentage points, but the effect is not statistically significant. This may have been expected since, from the perspective of the individual student, this tournament had a low probability of winning a small prize and targeted only top performance, thus potentially crowding out intrinsic motivation (Fryer 2013; Gneezy *et al.*, 2011).

"Hard" incentives, such as compulsion or assessment weights, in contrast, have a large positive effect on quiz participation. An assessment weight of 2.5% boosts quiz participation by 42 percentage points—a large increase given the low weight of the assessment. Doubling the incentive weight to 5% increases quiz participation by 62 percentage points, i.e., only about 1.5 times the effect of the lower weight. In comparison, declaring an assessment as compulsory

increases participation by 73 percentage points; as such an assessment weight of 5% is 85% as effective as compulsion. Therefore, assessment weights have a substantial effect on quiz participation.

We always reject the null hypothesis of (pairwise) equal parameter estimates for these three incentives at the 5% significance level (see bottom of Table 3), suggesting that effort increases with the assessment weight and is highest in a compulsory quiz. Our results are not sensitive to the fixed effects specification and are very similar when we use the full (unmatched) compared to the matched sample (see column 2 of Table 3). Importantly, the coefficient on the Cohort 2 dummy is not statistically significant, thus confirming the graphical evidence from Figure 1 that the two cohorts do not systematically differ in their intrinsic motivation to participate in quizzes. Additionally, we allow for an interaction term between incentives and cohort, and find no evidence that students from different cohorts react differently to the same incentives (see Table C2 in the appendix). Finally, ability is positively correlated with effort (Table 3, Column 3). Students in ability quartile 1 (respectively 2) are 7 (10) percentage points less likely to participate in a given quiz than top ability students. These estimates are significant at the 1% level. However, including ability controls does not alter the mean incentive effects on quiz participation. Quiz participation for students in the 3rd ability quartile does not significantly differ from the one of top ability students.

Student effort may also manifest itself in other dimensions, such as the intensity of participation (i.e., the time spent completing a quiz, normalized by the number of questions per quiz), and the productivity of effort which we measure through the quiz grade. Note that these variables are observed conditional on quiz participation, clearly creating selection between incentivized and non-incentivized weeks. Since, in the absence of incentives, more able/motivated students were more likely to participate, this selection biases our results towards zero. In spite of this bias, we still find statistically significant increases in intensity and

effectiveness of effort in weeks with hard incentives (see Table C.1 in the appendix): incentivized students spent 4 to 7 additional minutes per 10 quiz questions. The effort increase in weeks with 5% assessment weighting is comparable to that of a compulsory quiz. At a baseline of 25 minutes per quiz (and an average of 16.7 questions) in non-incentivized weeks, these effects are substantial. We find no increase in effort intensity from providing a soft incentive. Effort productivity also increases for the two strongest incentives, compulsion and 5% weighting, and we observe a small but statistically insignificant grade improvement with a 2.5% weight. Assessment weighting incentivizes increased effort in quizzes along all dimensions in a consistent manner.

In summary, the two cohorts are very similar in their intrinsic motivation to participate in (non-incentivized) quizzes and in their reaction to soft incentives. Neither tournament nor participation incentives are effective in encouraging effort, likely due to lack of salience, adverse effects on those who are unlikely to win, or a low probability of winning. However, even small assessment weights of 2.5 and 5% *increase* quiz participation *substantially*, by 42 to 62 percentage points, and also increase the effort put into quizzes.

3.2 Unintended Consequences? Effort displacement versus Effort Spillovers

When faced with a mixed schedule of incentivized and non-incentivized quizzes, students may simply shift effort between weeks rather than increase effort overall. Such displacement effects would represent unintended consequences of the new learning (and assessment) technique evaluated in this paper. Empirically, this would lead us to overestimate the impact of incentives on student effort. In this section we provide a set of direct tests for displacement effects across the weeks of our study.

Figure 1 and the statistically insignificant estimate for a cohort effect in Table 3 shows that participation in weeks without incentives is almost identical across the two groups, which is a first indication that the treatment group does not shift effort between weeks with and without incentives.

As our first test of displacement effects, we produce estimates of incentive effects *relative to participation in the first week of the year* (column 1 of Table 4). In the first week, fresher students did not know about the incentive structure of future quizzes. The scheme was announced only in the second week of term. Indeed, quiz participation in this week is not different between the two cohorts (t=0.52), so we use it as our baseline. Participation in week 1 was relatively high for a week without incentives (see Figure 1), so parameter estimates are marginally smaller than those in Table 3. Otherwise they are very similar. More importantly, participation in the first and subsequent non-incentivized weeks is not statistically different after controlling for a time trend. All results reject the displacement hypothesis and support our hypothesis that assessment weighting *increases student effort in incentivized weeks but does not displace it in other weeks*.

In a second test, we investigate whether an assessment in the previous week may induce lower quiz effort in the following week (column 2 of Table 4). We find little evidence that lagged assessment weighting incentives affect current quiz participation. The inclusion of lagged incentives does not reduce our parameter estimates for contemporaneous incentives (see column 1 in Table 3), and neither of the lagged terms is statistically significant. Additionally, we reject the null of joint significance of all lagged assessment weight (or compulsion) incentives; so again, we find no evidence of displacement effects.

Thirdly, we pursue an analogous strategy with (one-week) leads in incentives to assess whether an upcoming assessed quiz in period t+1 encourages students to participate in non-incentivized quizzes to practice in advance or obtain feedback. Alternatively, an upcoming quiz

could discourage effort in the week before. We find some evidence of anticipatory effects (in column 3): upcoming assessment weighted (at 2.5%) or compulsory quizzes increase present quiz effort. Contemporaneous incentive estimates do not change much when we include one-period forward dynamic effects, i.e., leads. Lead incentives for the strongly incentivized weeks (weight or compulsory) are jointly statistically significant. Hence, our estimates of the contemporaneous incentive effects appear to be underestimating the full effect of incentives: they affect participation in the week in which they are implemented but also have anticipatory or preparatory effects in the previous week. In summary, if anything, we find evidence of *positive effort* spillovers across weeks, but no evidence of discouragement of effort in non-incentivized weeks.

A final piece of evidence against the displacement hypothesis is provided through estimates of the impact of assessment weighting on the total number of quizzes Q submitted (see column 1 in Table 5). We sum up all weekly quizzes that were submitted before the in-term exam and estimate the model outlined in equation (3) in Section 2. Again, we find no evidence of displacement effects from assessment weighting; the overall number of submitted quizzes is significantly larger (1.2 additional quiz or 54%) in the treated group.

Figure 2 illustrates the shift in the distribution of the total number of completed quizzes. In particular we see a sharp drop in the fraction of students not completing any or just one quiz before the in-term assessment. For example, the proportion of students who never participate in quizzes decreases from 21 to 10%, while the proportion of students who always take quizzes increases from 1 to 7%. Additionally, we report a series of estimates on the probability of having completed at least *n* quizzes before the in-term exam (*n* takes values from 1 to 7) using linear probability models (see columns 2 to 8 of Table 5). We find increases in the number of submitted quizzes throughout the distribution of quiz effort (along the range of *n*), suggesting a lack of displacement effects not only at the mean but also across the distribution of effort.

Our estimate is the largest in the category "at least 2"; note that the median number of submitted quizzes in the control group is 2. We find a 56.5% increase in the probability of participating in 2 or more quizzes under assessment weighting. We also find large effects in the probability of having completed at least 3, 5 or 6 quizzes. Probability increases are largest close to the top, at 6 (or more) and 7 submitted quizzes, but this is due to the low frequency of students submitting all or almost all quizzes in the control group. In summary, we find that assessment weighting increases the probability of providing more effort at all levels, and that it particularly activates students with low effort levels (as can also be seen in Figure 2).

Our results show that assessment weighting strongly affects continuous learning effort by students, while other incentives are rather ineffective. We support this with lower bound estimates of incentive effects on the intensity and productivity in quizzes. We further show that there are no unintended consequences of providing assessment incentives in the form of displacement effects on non-incentivized effort across weeks. Total effort increases strongly by 54% on average.

3.3 Heterogeneous incentive effects on effort

We now relax the homogeneity assumption made in equation (1) to examine whether incentives have differential effects by gender or along the dimensions of ability and risk attitude. Solon et al. (2015) show that ignoring heterogeneous effects may provide inconsistent estimates of average effects due to differential subgroup population weights and subgroup-specific sampling variance. We restrict our analysis to cross-subgroup heterogeneity and assume a constant average treatment effect within subgroup. Recent studies incorporate both estimates of treatment effects across and within subgroups (Lehrer et al., 2016), and develop multiple testing strategies to test for these, while Bitler et al. (2016) develop a test under the

null that all heterogeneity is across subgroups. However, since our sub-group contain relatively few individuals we refrain from exploring within group heterogeneity.

We find no evidence of heterogeneous treatment effects by gender or risk attitudes (see columns 5 to 8 in Table 6). This is *a priori* surprising given the evidence in the literature that men respond more to competition (see Morin, 2015 for example) but might be driven by the design of the competition; rewarding only the best performer with a small prize. In further robustness checks, we do not find differential effects along the dimensions of age, subject of study, and patience. On the other hand, nationality and self-confidence matter: non-British students are more susceptible to incentives, as are self-confident students (see Table C.3 in the appendix).

We are most interested in heterogeneous effects by ability, as these would be indicative of whether assessment weighting (or other incentives) may help narrow the ability-performance gap by levelling effort. We construct ability quartiles based on university entry grades (for details, see Appendix A.1). There is no evidence of an impact of soft incentives (book voucher and access to additional study material) on quiz effort in any ability group (see columns 1 to 4 of Table 6). In contrast, heterogeneity matters for assessment weighting and compulsory quizzes; assessment incentives have a stronger impact among students in the lowest ability quartile and a lower impact in the top quartile. In weeks with assessment weighting (compulsion), lower ability students' participation in quizzes increases by about 30 (25) percentage points more than the participation of students in the top quartile. These differences are similar for both weights. One reason for the larger response of low-ability students to incentives is that they are 30% less likely to participate in quizzes in the absence of incentives than students with above the median ability. Thus, incentives provide stronger motivation to increase effort among lower ability students, reducing the participation gap between students.

3.4 Assessment Weighting Incentives and Student Performance

If cohort 2 students exert additional effort, i.e. participate in more quizzes, does student performance increase as well? We found that incentive-induced quiz effort does not reduce effort in non-incentivized quizzes and that assessment weighting induces a sizeable increase in total quiz effort. Still, incentives could lead to inter-temporal substitution of effort between term time and the exam preparation period, as such additional effort in quizzes may not improve performance. On the other hand, even incentives that are shifting effort may have beneficial effects: Even if they were neutral in terms of total effort, more continuous studying (during the term) may increase overall performance by enabling students to better follow lectures and seminars throughout the course. Finally, effort returns may vary depending on the effectiveness of the quiz in facilitating learning. For example, compulsory quizzes are good at increasing participation but might not be as effective a learning tool since, unlike assessed quizzes, they do not reward effort.

We will return to the question of displacement effects beyond the term in Section 3.5. In this section, we estimate the "net of potential displacement" combined effect of the effectiveness of quizzes and their impact on grades through increased effort. We present reduced form results of the introduction of assessment weighting on student performance based on equation (4). Of course, reduced form estimates could capture other performance factors that differ between the two cohorts. To test whether the difference in test scores is driven by the additional quizzes attempted by students from cohort 2, we extend the reduced form specification and include the number of quizzes attempted. If the difference in cohort performance is driven by quiz participation, we would expect that the cohort effect would become insignificant when we control for quiz participation. In addition, to test whether the incentives to participate in quizzes also affected the efficiency of quizzes in improving

knowledge, we also include interactions between cohort and the number of quizzes attempted (5).

We first measure short-run performance effects associated with the assessment of weekly quizzes. To prevent bias due to variation in exam difficulty or marking standards across years, our main estimates rely on grades from term-time tests. These tests consist of multiple-choice questions randomly drawn for each student from a large question bank, which did not change over time. The exams are graded automatically. As a consequence, our performance measure is, in expectation, identical across groups and not subject to marking bias. For each student, we observe two such exams, one per term. If assessments measure student performance with error, observing two exams reduces measurement error as long as the chance components of the two exams are uncorrelated. Grades are expressed in terms of standard deviations from the average exam grade of cohort 1 so that the average z-score for the first cohort has a mean of zero and a standard deviation of one. Exam participation does not substantially differ across cohorts, so our results are not driven by selection into the exam. Figure 3 shows that in both terms, the grade distribution at the in-term exam, our performance measure, shifts to the right for Cohort 2, which was induced to exert more effort via assessment weighting.

The introduction of assessment weighting increases exam performance by 0.27 of a standard deviation, and the estimated coefficient is statistically significant at the 5% level (see column 1 in Table 7). Next, we look beyond mean exam grades and estimate linear probability models for the probability of i) passing the exam (gaining at least 40%), and ii) obtaining an Upper Secondary or higher grade (at least 60%), (see columns 3 and 4 of Table 7). We find that while pass rates are not affected by assessment weighting (the baseline pass rate for these tests is 98%), there is an increase in the probability of obtaining an Upper Secondary grade or higher (by 11 percentage points, from a baseline of 73%). We further note that the variance of grades in these tests is significantly reduced by 17% ($F(267,331)= 1.30$).

In panel B of Table 7, we present our extended specification which provides a simple test that the reduced form estimates can be interpreted causally. When including the number of quizzes completed and its interaction with cohort, the cohort effect becomes insignificant; i.e. exam grades improve because students in cohort 2 completed more quizzes, not because of some other effect affecting cohort 2. Moreover, each completed quiz was as effective in producing grades in both cohorts; suggesting again that the two cohorts are similar and that the difference in exam performance is driven by the greater quiz completion rate of cohort 2. In summary, the two cohorts do not substantially differ in their grades or their returns to quiz effort other than through the impact of assessment weighting, which led cohort 2 to attempt more quizzes. Hence the increase in effort is productive (and quizzes as a learning tool are effective).

We also assess possible heterogeneity in the effect of assessment weighting on the test score by ability level. To do so, we interact ability quartile (based on A-levels or A-level equivalents) with the cohort indicator (Table 7, Column 2). The ability by cohort interactions are insignificant individually and jointly ($F(3,307)=1.41$); i.e. the increase in performance is observed for all ability groups. In Panel B, we include quiz completed and its interaction with cohort and the cohort effect disappears and only quiz completion has a significant correlation with test score. As such, we find consistent evidence that the improvement in grades observed for the cohort subjected to the assessment weighting of some of the weekly quizzes is correlated with quiz participation but not driven by cohort specific effects. While we recognize that quiz participation may be endogenously determined by ability and other non-cognitive skills, we find all our results to be qualitatively and quantitatively consistent. For example, if we multiply the average incentive-induced increase in the number of quizzes (1.266; see Table 5) with the effort return from one additional quiz of 0.165 (this is based on a specification which does not include the interaction term between quiz numbers and cohort, not reported in Tables), we

obtain an overall performance increase that is in the order of magnitude of the grade increase estimated in the reduced form specification.

We assess the possible non-linearity effect of quiz participation (see Table C4 in Appendix C). Note, in this specification we also find that cohort has no direct effect when controlling for quiz numbers and that the productive effect of quizzes on test score is not cohort dependent (F[7,307]=0.49) (column 2). Quiz participation has a significant impact on the in-term exam grade only when 4 (out of 7) or more quizzes have been completed during the term. There is no statistically significant difference in grades from completing 5, 6, or 7 quizzes. Since assessment weighting increased the proportion of students completing at least 4 tests by 40%, we would expect an average improvement on test scores of 0.25 of a standard deviation (0.40*0.634), yielding a test score effect that is almost identical to the effect estimated in the linear model (which amounts to 0.27 of a standard deviation); i.e. the average performance improvement on test score is consistent with the additional effort obtained from assessment weighting.

Overall, assessment weighting increased productive effort, leading to an overall grade improvement, mostly for students completing at least 4 or more quizzes per term. We find little evidence of heterogeneity by ability but report a relative reduction in grade dispersion (measured using the relative standard deviation) by 17%.

3.5 Inter-temporal and inter-subject substitution of effort?

In Section 3.2, we found no evidence of displacement of effort across weeks. However, additional effort during the term may facilitate learning (due to the cumulative pedagogic modularity of course contents or the effectiveness of quizzes as a learning tool); or it may simply lead to inter-temporal or inter-subject substitution of effort (Fryer and Holden, 2013).

We address these considerations in turn. First, if there were displacement effects, the return to (the same) quiz effort should be lower among students exposed to assessment weighting. However, we find no evidence of differential returns per quiz in Panel B of Table 7, and in Table C4 in the appendix.

Secondly, using the same specifications as for the in-term test grade, we find no evidence of an intertemporal shift away from effort outside the term which we measure through participation in non-incentivized revision quizzes that were made available during the exam preparation period (column 1 in panel A of Table 8).

Thirdly, we show that the performance increase under assessment weighting that we find for our preferred exam measure also applies to the end of year exam in "Principles" and the final course grade (see columns 2 and 3). These results should be viewed with caution, as the final course grade is directly affected by the additional marks students achieved via quizzes, and final exams difficulty and their marking may change between years. While the increase in the final exam grade is not statistically significant, the point estimate is very similar to the one observed for in-term tests. The overall course grade increases (statistically significantly) by 0.36 of a standard deviation. As for the in-term exam, we find no effect of assessment weighting on passing the course (column 4).

In the second panel of Table 8, we again report estimates from an extended model that includes the number of quizzes and their interaction with a cohort dummy. This is to test whether the reduced form estimates are driven by increased quiz participation or other cohort differences. Since cohort effects disappear when the number of quizzes is included, our results support the interpretation that the reduced form evidence can be interpreted as the effect of assessment weighting on final grades via its effect on quiz participation rather than as unobserved cohort differences such as differences in exam difficulty or grading bias.

Finally, we investigate inter-subject substitution of effort using students' average grade in three other courses that took place in the same study year. While the note of caution discussed above applies to the exam measures in these courses, variation in exam difficulty may be averaged out across these three compulsory courses. We find that assessment weighting is associated with a statistically significant and large increase in the mean grade in these courses, by 0.28 of a standard deviation. However, the reduced form effect is unlikely to only capture the effect of assessment weighting in "Principles". For example, unlike "Principles", these courses may have been subject to changes in course materials and course delivery across years. In consequence, even when we include quiz participation, the cohort effect remains (Panel B). This suggests that the reduced form effect is not solely capturing the increased quiz participation, and we should refrain from interpreting any grade improvement in these courses solely as a spill-over effect from the change in assessment weighting in "Principles".

3.6 Incentivized Learning

We now assess whether the increase in quiz participation due to assessment weighting affects longer-run performance. We do this by following "Principles" students into their second year and assessing whether they also perform better in two compulsory courses that have "Principles" as a pre-requisite. As Table 9 shows, we find positive parameter estimates at the mean and for greater scores (above 60%), statistically significant (marginally for the mean effect) only for Macroeconomics. The mean effect has the same size effect as in year 1, so long-run effects are potentially large. In the second panel, where we include the total number of quizzes in year 1 and its interaction with cohort, the cohort effect remains large although only marginally statistically significant or insignificant. Hence, the long-run effect might be driven by additional factors other than the assessment weighting in "Principles" in year 1.

4. Conclusions

A growing literature explores the role of incentives in fostering learning and improving education outcomes (see e.g. Lavecchia, Liu and Oreopoulos (2016) for a review). In this paper, we explore a set of possible incentive mechanisms. We find that tournament and pure participation incentives are largely ineffective in increasing study effort. In contrast, assessment weighting and compulsion are highly effective in inducing additional effort among students. Assessment weights need not be very high (and increasing them has diminishing marginal returns): in our study, weights of only 2.5% and 5% of the overall course grade trigger large participation increases of 45 and 62 percentage points, respectively. We also find that the effect of assessment weights varies with students' ability. High ability students display high participation rates in quizzes even in the absence of assessment weighting, so positive weighting is particularly effective in increasing effort among low and median ability students. Compulsion is even more effective at increasing quiz participation but might be more costly to scale up, due to administrative costs associated with checking the validity for excuses of non-response. Also, while it is effective at increasing participation, compulsion might not generate as much "productive effort" as assessment weighting, and thus have a lower impact on learning.

Recent criticisms of the use of incentives have warned against the danger of crowding out intrinsic motivation (see, e.g. Fryer (2011, 2013)). Students may either not respond to them at all or they may simply shift effort towards activities that are assessment weighted rather than increase effort overall. In this paper, we provide a detailed analysis of effort displacement effects. Effort is notoriously hard to measure as it manifests in various forms such as the level of endeavor in quizzes, lectures, seminars, self-study time, and varies in time and intensity. In this paper we present a broader set of effort measures than in previous studies and thus are able to shed new light on both the link between effort and performance and the potential

displacement of effort in various domains. All our results point in the same direction: students neither displace quiz effort between incentivized and non-incentivized weeks nor do we find evidence of effort displacement inter-temporally or across subjects. In contrast, we find some indication that there are positive spillovers on effort between weeks, performance in other courses in the same year, and marginally, on a related course a year later.

We conclude that assessment weights – or other mechanisms increasing quiz participation in general - are an effective means of improving students' performance: three assessed quizzes with a total weight of 10% of the final grade not only increased quiz participation but also grades on in-term tests by 0.27 of a standard deviation. These results suggest that there is no displacement of effort throughout the year due to the weekly incentives. We find some weak evidence of positive spillovers, but the set-up limits the causal interpretation of these estimates. The estimated effect on test-score is large, and similar to those found in studies that implemented costly financial incentives (see Angrist *et al.*, 2014, Garibaldi *et al.*, 2012 and Leuven *et al.* 2010). Our estimates are also comparable to the effects of relative and absolute feedback found in Bandiera *et al.* (2015). Note however, that the effects are non-linear and driven by getting students to complete at least 4 out of the 7 weekly quizzes, so there might be some limits on the efficiency of increasing the frequency of incentivized quizzes.

Recent work particularly tries to address performance deficits among lower ability students. These studies mostly focus on using relative achievement targets in the incentive design (Behrman *et al.*, 2015) or by targeting teacher performance (Figlio and Kenny, 2007). Due to its stronger effect on lower and median ability students, we suggest that assessment weighting seems to be an overlooked candidate in this debate: it helps reduce the within-group performance gap by about 17%. Future research is needed to investigate the extent to which:

lack of self-discipline in study behavior; strong time discounting; or lack of ability, are the cause of underperformance among low achieving students.

References

Angrist, J., Oreopoulos, P. and Williams, T. (2014) When Opportunity Knocks, Who Answers? New Evidence on College Achievement Awards, *Journal of Human Resources* 49(3), 572-610.

Angrist, J., Lang, D. and Oreopoulos, P. (2009) Incentives and Services for College Achievement: Evidence from a Randomized Trial, *American Economic Journal: Applied Economics* 1(1), 136-63.

Bandiera, O. Larcinese, V and Rasul, I. (2015) Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on Performance. *Labour Economics* 34, 13-25.

Barrow, L. and C. E. Rouse (2013) Financial Incentives and Educational Investment: The Impact of Performance-Based Scholarships on Student Time Use," NBER WP# 19351.

Behrman, E., Parker, S., Todd, P. and Wolpin, K. (2015) Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools, *Journal of Political Economy* 123(2), 325-364.

Bitler, M., Gelbach, J. and Hoynes, H. (2016) "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment," *Review of Economics and Statistics*. (Forthcoming)

Camerer, C. and Hogarth, R. (1999) The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework, *Journal of Risk and Uncertainty* 19,7-42.

Cason, T., Masters, W. and Sheremeta, R. (2010) Entry into winner-take-all and proportional-prize contests: An experimental study, *Journal of Public Economics* 94, 604-611.

Cha, P. and Patel, R. (2010) Rewarding Progress, Reducing Debt: Early Results from Ohio's Performance-Based Scholarship Demonstration for Low-Income Parents. MDRC, October 2010.

Dohmen, T., Falk, A., Huffman, A. and U. Sunde (2010) Are Risk Aversion and Impatience Related to Cognitive Ability? *American Economic Review* 100, 1238-1260.

Emerson, T. L. and Mencken, K. D. (2009). Homework: To require or not? Online graded homework and student achievement, *Perspectives on Economic Education Research* 7 (1), 20-42.

Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, *91*(5), 901-914.

Frey, B. and Jegen, R. (2002) Motivation crowding theory, *Journal of Economic Surv*eys 15(5), 589-623.

Fryer, R. (2011) Financial Incentives and Student Achievement: Evidence form Randomized Trials" *Quarterly Journal of Economics* 126, 1755-1798

Fryer, R. (2013) Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics,* 31, 373-427.

Fryer, R. And R. Holden (2012) Multitasking, Learning, and Incentives: A Cautionary Tale, NBER Working Paper 17752.

Garibaldi, P., Ichino, A., Giavazzi, F., and Rettore, E. (2012) College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities, *Review of Economics and Statistics* 94(3), 699-711.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011) When and Why Incentives (Don't) Work to Modify Behavior, *Journal of Economic Perspectives* 25(4), 191-210.

Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, *44*(2), 93-109.

Grove, W. A. and Wasserman, T. (2006) Incentives and student learning: A natural experiment with economics problems sets. *American Economic Review* 96(2), 437-41.

Koch, A., Nafziger, J., & Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior & Organization* 115(C), 3-17.

Lavecchia, A. M., Liu, H., & Oreopoulos, P. (2016). Behavioral economics of education: Progress and possibilities, In: E. A. Hanushek, S. Machin and L. Woessmann (eds), *Handbook of Economics of Education*, Volume 5, 1-74.

Lehrer, A., Pohl, V., and Song, K. (2016). Targeting Policies: Multiple Testing and Distributional Treatment Effects, *National Bureau of Economic Research*, WP22950Leuven, E., Osterbeek, H. and van der Klaauw, B. (2010) The effect of financial rewards on student's achievement: Evidence from a randomized experiment. *Journal of the European Economic Association* 8(6), 1243-65.

Morin, L.-P. 2015. "Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform," *Journal of Labor Economics* 33(2), 443-491.

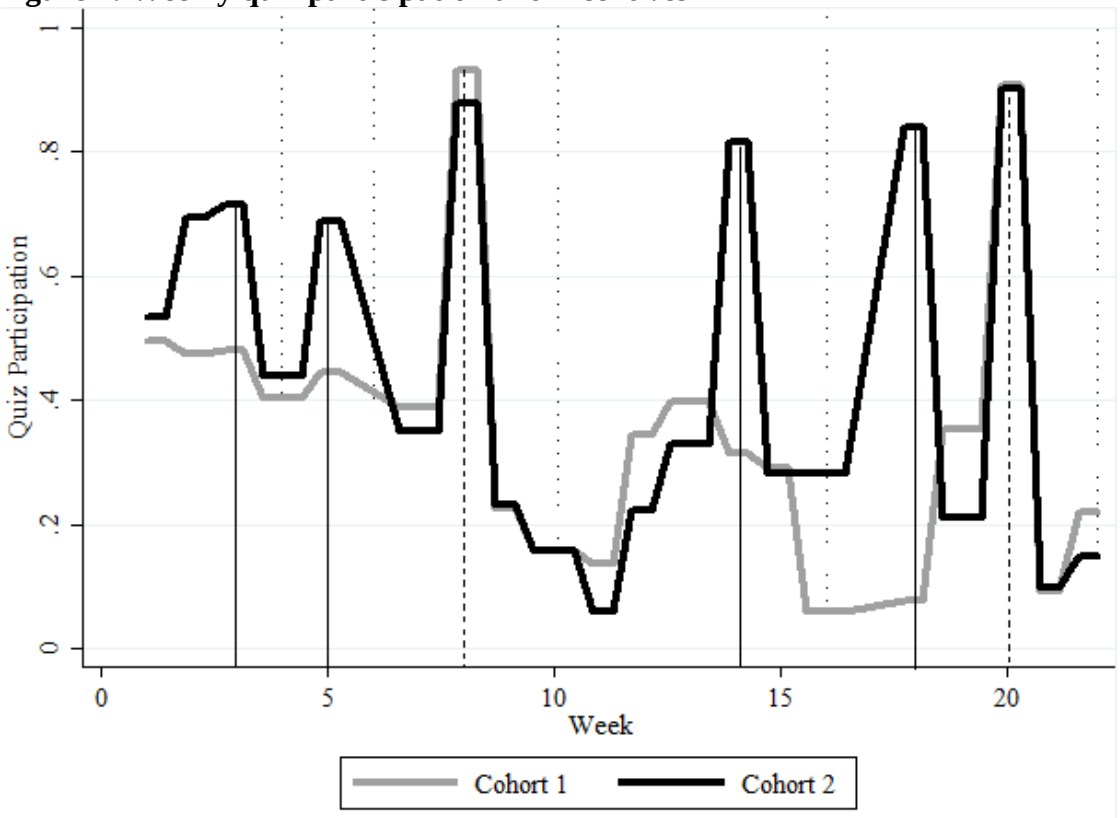
Pozo, S. and Stull, C.A. (2006) Requiring a Math skill unit: Results of a randomised experiment. *American Economic Review* 96 (2), 437-41.

Solon, G., Haider, S. and Wooldridge, J. (2015). "What Are We Weighting For?" *Journal of Human Resources* 50, pp. 301-316.

Trost, S., & Salehi-Isfahani, D. (2012). The effect of homework on exam performance: Experimental results from principles of economics. *Southern Economic Journal*, *79*(1), 224-242.

Tables and Figures

Figure 1: Weekly quiz participation and incentives



Source: Total number of students: 206 in Cohort 1 and 240 in Cohort 2
Note: Solid vertical lines refer to Assessment 2.5% (Cohort 2 only)
Dashed vertical lines refer to compulsory quiz (Cohort 1) or Assessment 5% (Cohort 2)
Dotted vertical lines refer to soft incentives: book voucher and solution provision

Figure 2: Total number of quizzes completed before in-Term Exam

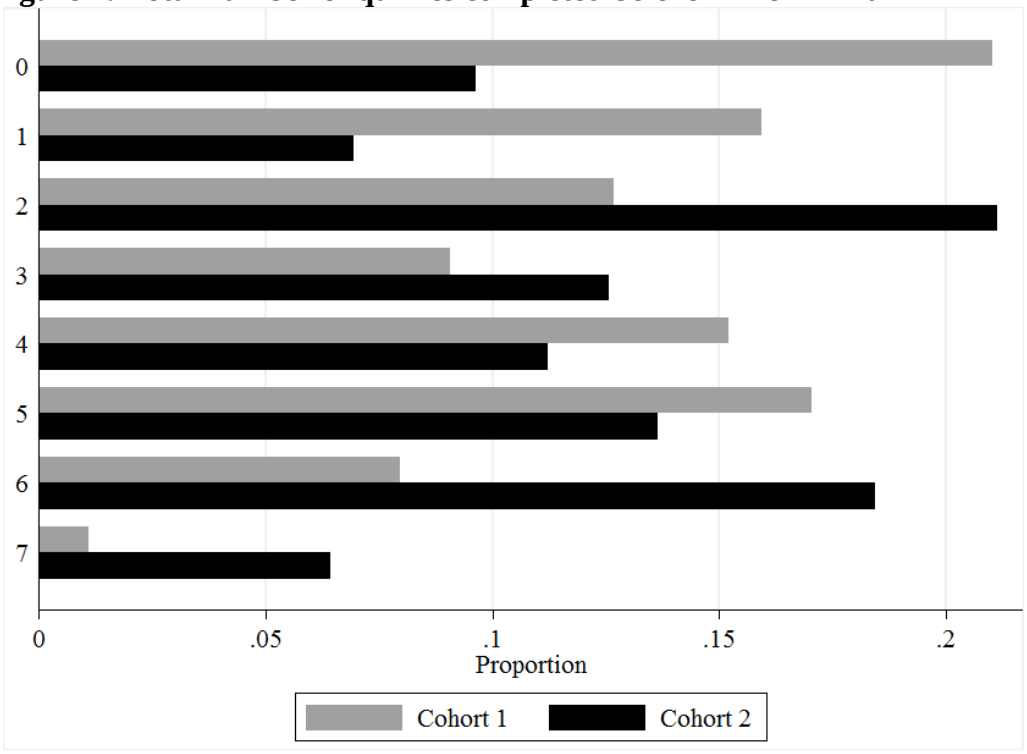


Figure 3: Distribution of In-Term Exam grades

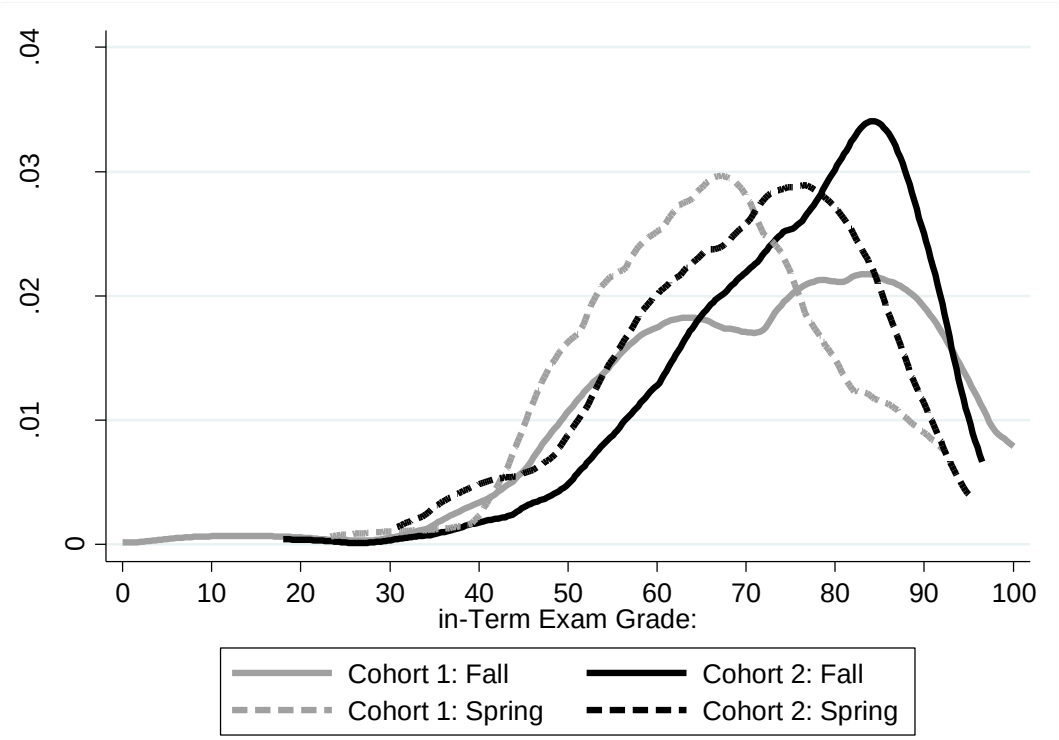


Table 1: Timing of incentives by cohort

	Fall		Spring	
Week	**Cohort 1**	**Cohort 2**	**Cohort 1**	**Cohort 2**
1	O	O	O	O
2	O	O	O	O
3	O	Assessed 2.5%	O	Assessed 2.5%
4	Solution	Solution	O	O
5	O	Assessed 2.5%	Voucher	Voucher
6	Term break		Term break	
7	O	O	O	Assessed 2.5%
8	Compulsory	Assessed 5%	O	O
9	O	O	Compulsory	Assessed 5%
10	Voucher	Voucher	O	O
11	O	O	Solution*	Solution*

*N*ote: "O": designates weeks without incentives to complete the online quiz; quiz is pure formative feedback; "Solution": Access to the weekly exercise sheet solutions conditional on quiz participation; * indicates that access to solution was conditional on getting a mark of 30 or above.
"Voucher": £20 book voucher - prize for best quiz performance; "Assessed 2.5%" ("Assessed 5%"): Assessed quiz, counting 2.5% (5%) towards the overall course grade; "Compulsory": Quiz mandatory part of 4 pieces of coursework (3 out of 4 are required)

Table 2: Descriptive statistics

	Full sample		Matched sample	
	Cohort 1	**Cohort 2**	**Cohort 1**	**Cohort 2**
Panel A. Student characteristics				
Age (in months)	233.6 (16.60)	230.5** (13.70)	227.2 (9.812)	229.1* (11.48)
Male	0.529 (0.500)	0.637** (0.482)	0.643 (0.481)	0.615 (0.488)
UK citizen	0.294 (0.457)	0.538*** (0.500)	0.582 (0.495)	0.594 (0.492)
Chinese citizen	0.127 (0.334)	0.081 (0.273)	0.069 (0.254)	0.059 (0.236)
Econ Major	0.485 (0.501)	0.709*** (0.455)	0.742 (0.439)	0.717 (0.452)
Ability[a]	330.7 (83.81)	303.7*** (66.38)	312.41 (72.83)	306.47 (61.09)
No. Obs.	204	234	138	187
Panel B. Student preference parameters and (usually unobserved) characteristics				
Confidence	12.986 (2.710)	12.642 (2.606)	12.736 (2.801)	12.632 (2.695)
Patience	4.349 (5.307)	4.387 (5.363)	4.356 (5.394)	4.032 (5.370)
Risk attitudes	6.319 (1.451)	6.407 (1.310)	6.295 (1.404)	6.429 (1.253)
No Obs.	142	155	103	128
Panel C. Student effort and performance				
No. quizzes attempted	7.838 (4.757)	9.884*** (5.171)	7.650 (7.650)	9.914*** (5.356)
Quiz grade (%)[1]	56.68 (10.06)	64.54*** (9.30)	59.16 (9.667)	65.19*** (9.205)
Quiz duration (mins)[1]	33.406 (9.62)	34.845 (9.50)	32.764 (8.564)	35.305***(9.651)
Self-study[1] (hrs per week)	2.839 (1.532)	4.469***(2.531)	2.788 (1.443)	4.451*** (2.568)
Lecture attendance[1]	0.907 (0.185)	0.902 (0.172)	0.940 (0.137)	0.897** (0.180)
In-term exam grade	68.13 (12.72)	72.27*** (10.80)	69.66 (12.16)	73.16*** (10.50)
Final exam grade[b]	46.38 (15.50)	50.03** (16.65)	48.92 (14.24)	52.11** (15.85)
No Obs.	204	234	136	183

*N*ote: ***/** indicate significant mean differences between waves at the 1 or 5% significance level. Standard deviations reported in parentheses. Matched samples obtained from kernel matching (Epanechnikov) with bandwidth (0.01).

[a]: Ability not observed for all participants, sample sizes in the full sample are 145 (202) for Cohort 1(2).

[b]: Final grades are observed for 193 (231) students only for Cohort 1 (2).

[1]: observed conditional on quiz participation only.

Econ Major denotes Economics or Economics and Finance students.

Variables in Panel B are measured in week 1 of term 1 (see appendix A.1): *Risk attitudes* is the sum of scores obtained from the questions: Do you have a personal medical insurance? Do you smoke? Do you take out travel insurance? Have you incurred interest charges on your credit cards? Do you play lotteries? Do you have a savings account? Did you play slot machines last week? Do you go out of your way to cross the road at pedestrian crossings? Do you do any dangerous sport? *Confidence* is the sum of scores obtained from answers to statements: I feel comfortable speaking to a bank manager about loans, I enjoy challenging situations, I'm not scared of being in debt, I handle uncertainty well. *Patience* is elicited from 5 questions regarding the present values of hypothetical prizes one year later.

Table 3: Impact of Incentives on Student Effort (measured as quiz participation)

	Matched sample FE	All OLS	Ability Sample
Solution	0.019 (0.032)	0.013 (0.030)	0.018 (0.031)
Book voucher	-0.087 (0.069)	-0.091 (0.067)	-0.084 (0.066)
Assess 2.5%	0.420 (0.084)***	0.413 (0.079)***	0.411 (0.082)***
Assess 5%	0.622 (0.036)***	0.628 (0.036)***	0.623 (0.039)***
Compulsory	0.727 (0.041)***	0.693 (0.036)***	0.686 (0.039)***
Cohort 2		0.001 (0.029)	0.014 (0.027)
Ability Q1			-0.074 (0.015)***
Ability Q2			-0.098 (0.013)***
Ability Q3			0.026 (0.012)**
Observations [individuals]	6500 [325]	8480 [424]	6880 [424]
Test for equality of incentives [a]			
H_0: $\beta_{Ass2.5} = \beta_{Ass5}$	F(1,39)=5.0**	F(1,39)=6.4**	F(1,39)=5.5**
H_0: $\beta_{Ass5} = \beta_{Comp.}$	F(1,39)=9.97***	F(1,39)=4.6**	F(1,39)=3.8

*N*ote: ***/** indicate statistical significance from zero at the 1 or 5% significance level.
Estimates are based on the matched sample. Robust standard errors are clustered at the cohort/week level.
Other independent variables are: an indicator of term, a term-specific time trend, dummies for assessments and essays in that week in other modules. In the OLS specification, we include the time-invariant individual characteristics – a linear term in age (in months), sex, dummies for Chinese Nationality and other non-UK nationalities, and degree subject. Robust standard errors are clustered at cohort/week level
[a]: F-tests at the bottom of the table refer to the Null hypothesis of equal parameter estimates i) Assess 2.5% = Assess 5%, ii) Assess 5% = compulsion.

Table 4: Displacement Effects of Incentives on Student Effort Across Weeks (measured as quiz participation)

	FE (rel. to first week)	FE + lagged incentive[b]	FE + forward incentive[b]
No incentives	-0.039 (0.071)		
Contemporaneous incentives in t			
Solution	-0.015 (0.050)	0.030 (0.052)	-0.039 (0.050)
Book voucher	-0.127 (0.0103)	-0.088 (0.099)	-0.134 (0.060)**
Assess 2.5%	0.387 (0.090)***	0.415 (0.097)***	0.463 (0.068)***
Assess 5%	0.590 (0.067)***	0.622 (0.096)***	0.601 (0.052)***
Compulsory	0.689 (0.080)***	0.754 (0.083)***	0.698 (0.065)***
Lags (t-1) respectively leads(t+1) of incentives			
Solution		-0.070 (0.129)	-0.093 (0.084)
Voucher		-0.010 (0.087)	-0.014 (0.048)
Assessed 2.5%		-0.037 (0.052)	0.114 (0.045)**
Assessed 5%		-0.016 (0.116)	0.025 (0.059)
Compulsory		-0.008 (0.101)	0.214 (0.063)***
Observations	6500	5525	5525
[individuals]	[325]	[325]	[325]
Test for equality of incentives and joint significance of lags (respectively leads):[a]			
H_0: $\beta_{Ass2.5} = \beta_{Ass5}$	F(1,33)5.0**	F(1,33)=3.4	F(1,33)=2.4
H_0: $\beta_{Ass5} = \beta_{Comp.}$	F(1,33)= 9.2***	F(1,33)= 5.4**	F(1,33)= 8.5***
H_0: $\beta_{lag\ /\ lead} = 0$		F(3,33)= 0.2	F(3,33)=25.9***

*N*ote: ***/** indicate statistical significance from zero at the 1 or 5% significance level.
Estimates are based on the matched sample. Robust standard errors are clustered at the cohort/week level. Other independent variables are: an indicator of term, a term-specific time trend, dummies for assessments and essays in that week in other modules. "No incentive" refers to quizzes in weeks without incentives (except for the first quiz). [a]: F-tests are for joint significance of lag (forward) coefficients on assessed quizzes (2.5 and 5%) and compulsion. [b]: We exclude the first week of each term in columns 2 and 3 to account for holiday breaks.

Table 5: Displacement vs Enhancement Effects in the Number of Submitted Quizzes before the in-term Exam

	Number of Submitted Quizzes before the in-term Exam							
	Nbr quizzes	**At least 1**	**At least 2**	**At least 3**	**At least 4**	**At least 5**	**At least 6**	**At least 7**
Cohort 2	1.266***	0.159***	0.303***	0.206***	0.148**	0.192***	0.199***	0.058***
	(0.284)	(0.048)	(0.061)	(0.066)	(0.070)	(0.052)	(0.031)	(0.13)
Mean Cohort 1	2.340	0.749	0.536	0.431	0.361	0.202	0.055	0.006
Mean size effect (in %)	54.1	21.2	56.5	47.8	40.1	95.0	361.8	966.7

*N*ote: Matched Sample (n= 650 observations; N=325 students);
Robust standard errors clustered at the individual level. ***/** indicate statistical significance from zero at the 1 or 5% significance level.
Additional covariates included in the regressions are cohort, gender, age, nationality, subject of degree, and dummies for ability quartile (based on UCAS score)

Table 6: Heterogeneous incentive effects on student effort (relative to the first week)

	Ability				Gender		Risk attitude	
	Q1	**Q2**	**Q3**	**Q4**	**Female**	**Male**	**Low**	**High**
No incentive	0.065	-0.104	-0.062	-0.060	-0.007	-0.058	-0.020	-0.014
	(0.083)	(0.088)	(0.108)	(0.103)	(0.080)	(0.072)	(0.095)	(0.063)
Solution	0.043	-0.045	-0.018	-0.054	-0.003	-0.022	0.005	-0.016
	(0.087)	(0.074)	(0.078)	(0.106)	(0.072)	(0.053)	(0.074)	(0.056)
Book voucher	0.001	-0.169	-0.185	-0.205	-0.117	-0.132	-0.108	-0.104
	(0.095)	(0.131)	(0.127)	(0.132)	(0.112)	(0.103)	(0.133)	(0.099)
Assess 2.5%	0.500***	0.384***	0.312***	0.205	0.364***	0.400***	0.377***	0.341***
	(0.110)	(0.085)	(0.115)	(0.138)	(0.109)	(0.085)	(0.116)	(0.104)
Assess 5%	0.679***	0.575***	0.564***	0.389***	0.606***	0.574***	0.595***	0.554***
	(0.083)	(0.085)	(0.093)	(0.110)	(0.078)	(0.069)	(0.093)	(0.065)
Compulsory	0.865***	0.631***	0.623***	0.589***	0.697***	0.684***	0.648***	0.740***
	(0.090)	(0.105)	(0.112)	(0.116)	(0.091)	(0.082)	(0.104)	(0.076)
Constant	0.446	0.548	0.700	0.710	0.682	0.507	0.669	0.580
	(0.055)	(0.020)	(0.030)	(0.092)	(0.030)	(0.021)	(0.022)	(0.028)
Observations	1860	2140	1380	1120	2820	3680	2020	2340

*N*ote:

Estimates are based on the matched sample. Robust standard errors clustered at the cohort/week level. ***/** indicate statistical significance from zero at the 1 or 5% significance level.

Other independent variables are: an indicator of term, a term-specific linear time trend and a dummy for assessments (essays) due in the same week in other modules.

Table 7: Impact of Student Effort on Normalized In-Term Exam Grades

Outcome	**Mean Normalized Grade**		**Pass (>=40%)**	**Upper secondary or above (>=60%)**
Population	***All***	***All Heterogenous ability effect***	***All***	***All***
Panel A				
Cohort 2	0.268**	0.400**	0.008	0.110***
	(0.131)	(0.203)	(0.012)	(0.056)
Ability Q1 * Cohort 2		-0.074 (0.253)		
Ability Q2 * Cohort 2		-0.467 (0.372)		
Ability Q3 * Cohort 2		0.235 (0.279)		
Panel B				
Number of Quizzes	0.204***	0.216***	0.003	0.063***
	(0.04751)	(0.047)	(0.005)	(0.020)
Cohort 2	0.09981	0.302	-0.012	0.037
	(0.17936)	(0.257)	(0.020)	(0.104)
Nbr Quizzes * Cohort	-0.0391	-0.064	0.004	-0.009
	(0.05046)	(0.048)	(0.007)	(0.024)
Ability Q1 * Cohort 2		-0.226 (0.249)		
Ability Q2 * Cohort 2		-0.382 (0.285)		
Ability Q3 * Cohort 2		0.333 (0.238)		
Observations	600	600	600	600

*N*ote:
Estimates are based on the matched sample (n= 600 observations ; N=308 students).
Standard errors are clustered at the individual level. ***/** indicate statistical significance from zero at the 1 or 5% significance level.
Controls include: dummies for gender, Chinese nationals and other non-UK nationals, subject of degree and term, a linear trend in age (in month) and ability quartiles.

Table 8: Displacement of Effort and Learning – First Year Outcomes

Outcome	Attempted revision quizzes	Final Exam grade	Course grade	Pass Course	Mean grade in other courses
Panel A					
Cohort 2	0.057 (0.076)	0.265 (0.171)	0.359** (0.169)	0.057 (0.052)	0.281*** (0.096)
Panel B					
Number of Quizzes	0.059*** (0.008)	0.114*** (0.027)	0.115*** (0.027)	0.022*** (0.008)	0.067*** (0.010)
Cohort 2	0.121 (0.085)	0.328 (0.255)	0.194 (0.240)	-0.008 (0.120)	0.424** (0.165)
Nbr Quizzes * Cohort	-0.020** (0.010)	-0.034 (0.029)	-0.011 (0.027)	0.001 (0.010)	-0.030** (0.014)
Observations	325	325	325	325	325

*N*ote: Estimates are based on the matched sample.
***/** indicate statistical significance from zero at the 1 or 5% significance level.
Estimates are based on the matched sample. Controls include: dummies for gender, Chinese Nationals, other non-UK nationals, and a linear in age (in month), ability quartiles.

Table 9: Learning - Long-run Outcomes

	Second year grade in: *Macroeconomics*			*Microeconomics*		
	Mean Normalised Grade	*Grade >=40*	*Grade >=60*	*Mean Normalised Grade*	*Grade >=40*	*Grade >=60*
Cohort 2	0.284*	-0.107	0.231***	0.131	0.099	-0.007
	(0.154)	(0.062)	(0.057)	(0.205)	(0.080)	(0.085)
Nbr Quizzes in year 1	0.052***	0.003	0.019*	0.118***	0.034***	0.038***
	(0.018)	(0.008)	(0.011)	(0.028)	(0.012)	(0.013)
Cohort2	0.208	-0.148	0.208*	0.617*	0.297*	0.102
	(0.258)	(0.108)	(0.112)	(0.338)	(0.172)	(0.139)
Nbr Quizzes * Cohort 2	-0.003	0.003	-0.002	-0.071**	-0.026*	-0.018
	(0.023)	(0.010)	(0.012)	(0.031)	(0.014)	(0.014)
Observations	272			272		

*N*ote: Estimates are based on the matched sample.
***/**/* indicate statistical significance from zero at the 1, 5 or 10% significance level.
Controls include: dummies for gender, Chinese Nationals, other non-UK nationals, and a linear trend in age (in month) and ability quartiles.

Appendix A: Description of data

Appendix A.1: Variable descriptions

Age: We report age in months rather than years to allow for performance differences by month of birth which have been shown to have a substantial effect on educational performance (see Crawford et al. 2014).

Ability: We proxy ability through UCAS tariff scores, i.e. school leaving grades, that were obtained before the start of the first year and the natural experiment. For international students, we use a combination of the academic equivalencies scales published by the University of Brighton (www.brighton.ac.uk/international/equivalencies) and the scales used by the admissions office of the College in which we conduct the experiment. We deviate in the valuation of the international baccalaureat as the equivalence scales seem too conservative given the high quality of this school degree programme.

Survey-based measures:

The following variables were obtained from a survey that was run each year at the beginning of the year in-class. Response rates for these survey are 74% (67%) for cohort 1 (cohort 2).

Risk attitudes: These are obtained as the sum of answers on the questions: Do you have a personal medical insurance? Do you smoke? Do you take out travel insurance? Have you incurred interest charges on your credit cards? Do you play lotteries? Do you have a savings account? Did you play slot machines last week? Do you go out of your way to cross the road at pedestrian crossings? Do you do any dangerous sport? Answer options were `yes' or `no'. Higher numbers denote more risk tolerance.

Patience: These are elicited from five questions regarding the present values of hypothetical prizes in one year's time. Respondents state whether they feel better off, worse off or consider the current and future payments equally valuable. Patience is then scored based on their answers regarding five payment pairs and lies between -2.5 (very impatient) and 12.5 (very patient).

Self-Confidence: Confidence is the sum of answers obtained from answers to statements: I feel comfortable speaking to a bank manager about loans, I enjoy challenging situations, I'm not scared of being in debt, I handle uncertainty well. All questions are scored on a Likert scale from 1 to 5. Higher values denote more self-confidence.

Appendix A.2: Dropout rates (Sample Attrition)

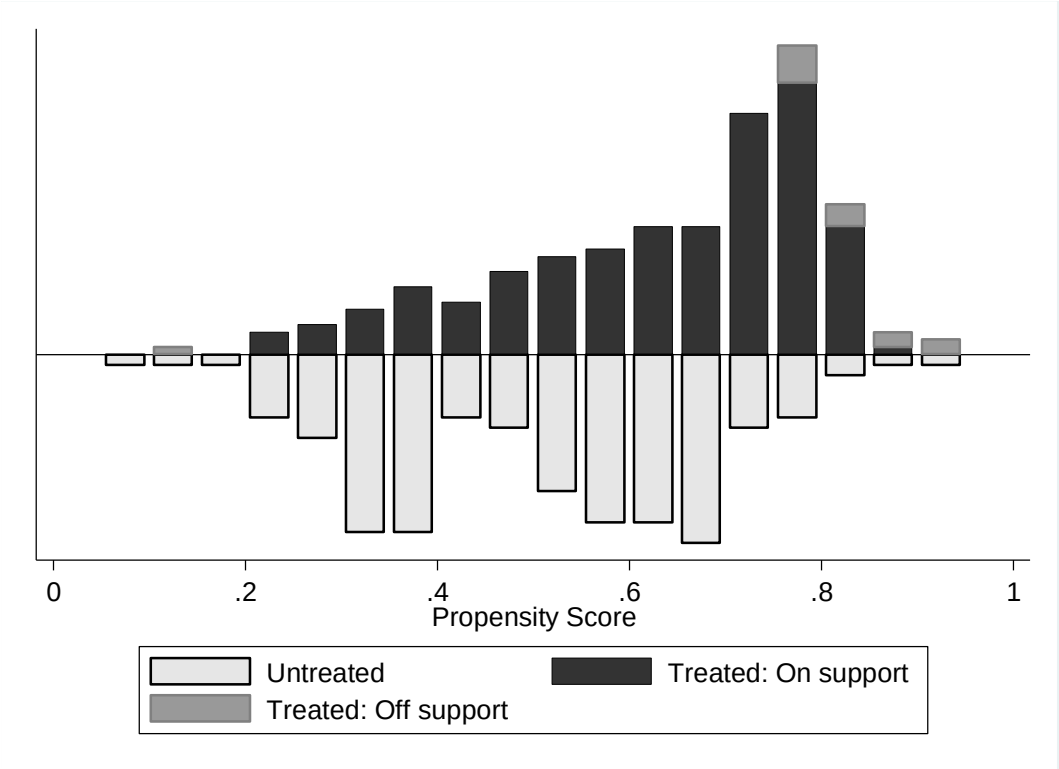
Drop out is very low at this institution, only 14 students across the two cohorts do not have a reported final grade. This includes dropout students as well as visiting students whose visit lasted for only one term. On average, these students participated in 3.9 quizzes, and 3 of them did not participate in any. These 14 participants do not impact on our main results as they cannot be matched and drop out of the analysis. Otherwise, there is no source of attrition as we use administrative student data.

Appendix B: Matching procedure

The matching is obtained using an Epanechnikov kernel with bandwidth 0.01 based on a program developed by Leuven and Sianesi. The propensity score is estimated by including the following variables: age, gender, nationality, major subject and ability (UCAS score);16 individuals from the second cohort are not matched and are dropped from the analysis.

Figure B.1 shows the Propensity score distribution and the overlapping support.

Figure B.1: Distribution of Propensity Score by Support Status



Appendix C: Robustness checks

Appendix C.1: Additional evidence on the impact of incentives on student effort

Table C.1 shows that assessment weighting increases effort also in terms of the intensity of effort, i.e. time spent doing a quiz, and improves quiz performance (statistically significantly only for the 5% assessment weighting). The quantitative effect of 5% assessment weighting is comparable to that of a compulsory quiz in terms of effort intensity and not much larger than for the lower weight of 2.5%. Table C. 2 shows both cohorts reacted identically to the joint incentives they faced – the soft incentives, compulsion (as well as non-incentivized weeks). In Table C.3, we find no heterogeneity in the reaction to incentives between economics and non-economics majors, by age or patience. There are heterogeneous effects by nationality: British students react strongest to incentives, illustrated by significantly lower participation in non-incentivized quizzes and increased participation in incentivized ones relative to non-Chinese foreign students, our base group. Finally, we find that confident students participate more in assessment weighted quizzes.

Table C.1: Impact of Incentives on Student Effort, alternative measures

	Log Time spent per question	Normalised Grade
Book voucher	0.009 (0.128)	0.068 (0.332)
Solution	-0.006 (0.108)	-0.211 (0.159)
Assess 2.5%	0.529 (0.149)***	0.072 (0.144)
Assess 5%	0.370 (0.141)**	0.642 (0.194)***
Compulsory	0.311 (0.102)***	1.089 (0.206)***
Observations [individuals]	2717 [317]	2723 [317]
R^2	0.26	0.09

*N*ote: Estimates based on the matched sample and using individual fixed effects. Robust standard errors clustered at cohort/week level. ***/**/* indicate statistical significance from zero at the 1, 5 or 10% significance level. Other independent variables are: an indicator of term, term-specific trend, gender, Chinese National, other non-UK national dummies, subject of degree and a linear term in age (in month), dummies for assessments in that week in other modules, essay in that week in other modules.

Table C.2: Impact of Incentives on Student Effort for both Cohorts

	Quiz participation	Quiz participation
No incentives	-0.021 (0.065)	-0.015 (0.062)
No incentive * Cohort 2		-0.014 (0.039)
Soft incentives	-0.047 (0.061)	-0.060 (0.075)
Soft incentives * Cohort 2		0.024 (0.054)
Assess 2.5%	0.399 (0.092) ***	0.397 (0.094) ***
Assess 5%	0.622 (0.056) ***	0.619 (0.062) ***
Compulsory	0.727 (0.066) ***	0.728 (0.061) ***
Observations [individuals]	6500 [325]	6500 [325]
H_0: $\beta_{none*coh2} = \beta_{soft*coh2=0}$		F(2,39)=0.16

*N*ote: Estimates based on the matched sample and using individual fixed effects. Robust standard errors clustered at cohort/week level. ***/**/* indicate statistical significance from zero at the 1, 5 or 10% significance level. Other independent variables are: an indicator of term, a term-specific time trend, dummies for gender, Chinese Nationality and other non-UK nationalities, subject of degree and a linear term in age (in months), dummies for assessments and essays in that week in other modules. "No incentive" refers to quizzes in weeks without incentives (except for the first quiz).

Table C.3: Heterogeneous incentive effects on student effort

	Nationality		Major		Age		Self-confidence		Patience	
	British	Non-British	Econo-mics	Other	Below median	Above median	Low	High	Low	High
No incentive	-0.106	0.056	-0.059	0.014	-0.090	0.049	-0.079	-0.011	-0.023	-0.033
	(0.092)	(0.088)	(0.078)	(0.065)	(0.086)	(0.077)	(0.090)	(0.066)	(0.080)	(0.070)
Solution	-0.082	0.081	-0.026	0.015	-0.047	0.041	-0.056	0.015	-0.009	-0.014
	(0.071)	(0.085)	(0.057)	(0.059)	(0.061)	(0.074)	(0.074)	(0.052)	(0.064)	(0.054)
Book	-0.201	-0.021	-0.134	-0.107	-0.185	-0.027	-0.191	-0.082	-0.119	-0.111
voucher	(0.129)	(0.0105)	(0.105)	(0.113)	(0.119)	(0.098)	(0.119)	(0.099)	(0.113)	(0.109)
Assess 2.5%	0.309***	0.497***	0.395***	0.367***	0.354***	0.445***	0.322***	0.433***	0.329***	0.393***
	(0.103)	(0.113)	(0.084)	(0.115)	(0.102)	(0.096)	(0.114)	(0.079)	(0.094)	(0.087)
Assess 5%	0.497***	0.712***	0.573***	0.620***	0.544***	0.659***	0.498***	0.650***	0.558***	0.582***
	(0.090)	(0.091)	(0.072)	(0.071)	(0.081)	(0.082)	(0.086)	(0.066)	(0.081)	(0.068)
Compulsory	0.605***	0.808***	0.708***	0.631***	0.615***	0.821***	0.611***	0.743***	0.686***	0.688***
	(0.105)	(0.093)	(0.085)	(0.079)	(0.094)	(0.085)	(0.099)	(0.077)	(0.068)	(0.080)
Constant	0.630	0.490	0.557	0.614	0.643	0.449	0.656	0.513	0.614	0.632
	(0.044)	(0.064)	(0.011)	(0.013)	(0.028)	(0.054)	(0.028)	(0.019)	(0.030)	(0.014)
Observations	3380	3120	4100	2400	3820	2680	2880	3620	2400	2100

*N*ote: Weights obtained from propensity score matching. Robust standard errors clustered at the cohort/week level. ***/**/* indicate statistical significance from zero at the 1, 5 or 10% significance level.

Other independent variables are: an indicator of term, a term-specific linear time trend and a dummy for assessments (essays) due in the same week in other modules.

Appendix C.2: Additional evidence on the impact of effort on student performance

C.2.1 Non-linear effort effects

Finally, we allow for nonlinear effort effects on performance using dummies for the number of quizzes (Table C.4). In column 1, we find a discontinuous jump in exam performance at 4, and a smaller increase at 5 quizzes (relative to never taking a quiz). Higher effort, i.e. doing 6 or 7 quizzes does not result in a significant increase in performance relative to submitting 5 quizzes. Again, these estimates are robust to allowing for cohort-specific effort returns as column 2 shows. Furthermore, while we find evidence for the joint significance of the (non-linear) quiz dummies, we do not find evidence that they are statistically significantly different across cohorts (F-statistic- 0.49).

Table C.4: Impact of Student Effort on Normalized In-Term Exam Grades.

	OLS	**OLS**
Nbr Quizzes		
Cohort 2	0.003 (0.105)	0.246 (0.269)
1 Quiz	-0.095 (0.280)	-0.085 (0.326)
2 Quizzes	0.267 (0.208)	0.187 (0.263)
3 Quizzes	0.352 (0.223)	0.436 (0.246)*
4 Quizzes	0.634 (0.278)**	0.672 (0.345)*
5 Quizzes	0.824 (0.239)***	0.872 (0.255)***
6 Quizzes	0.833 (0.261)***	0.742 (0.320)**
7 Quizzes	0.842 (0.228)***	0.981 (0.340)***
Quiz * cohort 2		Yes
F-test		
quiz dummies F(7,307)	7.82***	5.08***
across cohorts F(7,307)		0.49

Note: Other controls as in Table 7. Robust standard errors adjusted for clustering at the individual level.
quiz dummies F(7,307) reports the results from a test of joint significance of the number of quizzes completed variables.

Across cohorts F(7,307) reports the results from a test of joint significance of the interaction terms between number of quizzes and cohort.

Additional References

Crawford, C., Dearden, L. and Greaves, E. (2014) The drivers of month-of-birth differences in children's cognitive and non-cognitive skills, *Journal of the Royal Statistical Society: Series A* 177, 829-860.

Leuven, E., and Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Downloadable from http://ideas.repec.org/c/boc/bocode/s432001.html.